Proceedings Paper:
Ding, Z., Wang, A., Chen, H. et al. (5 more authors) (2023) Exploring structured semantic prior for multi label recognition with incomplete labels. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Proceedings. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 17-24 Jun 2023, Vancouver, BC, Canada. Institute of Electrical and Electronics Engineers , pp. 3398-3407. ISBN 9798350301304

https://doi.org/10.1109/CVPR52729.2023.00331

Exploring Structured Semantic Prior for Multi Label Recognition with Incomplete Labels

Zixuan Ding[1,4*] Ao Wang[2,3,4*] Hui Chen[2,3,†] Qiang Zhang[1]
Pengzhang Liu[5] Yongjun Bao[5] Weipeng Yan[5] Jungong Han[6,7]
[1]Xidian University [2]Tsinghua University [3]BNRist
[4]Hangzhou Zhuoxi Institute of Brain and Intelligence [5]JD.com
[6]Department of Computer Science, the University of Sheffield, UK
[7]Centre for Machine Intelligence, the University of Sheffield, UK
dingzixuan@stu.xidian.edu.cn wa22@mails.tsinghua.edu.cn qzhang@xidian.edu.cn
{jichenhui2012,jungonghan77}@gmail.com {Paul.yan, baoyongjun, liupengzhang}@jd.com

Abstract

Multi-label recognition (MLR) with incomplete labels is very challenging. Recent works strive to explore the image-to-label correspondence in the vision-language model, i.e., *CLIP [22], to compensate for insufficient annotations. In spite of promising performance, they generally overlook the valuable prior about the label-to-label correspondence. In this paper, we advocate remedying the deficiency of label supervision for the MLR with incomplete labels by deriving a structured semantic prior about the label-to-label correspondence via a semantic prior prompter. We then present a novel Semantic Correspondence Prompt Network (SCPNet), which can thoroughly explore the structured semantic prior. A Prior-Enhanced Self-Supervised Learning method is further introduced to enhance the use of the prior. Comprehensive experiments and analyses on several widely used benchmark datasets show that our method significantly outperforms existing methods on all datasets, well demonstrating the effectiveness and the superiority of our method. Our code will be available at* https://github.com/jameslahm/SCPNet.

1. Introduction

Multi-label recognition (MLR) aims to describe the image content with various semantic labels [5, 26, 29, 30]. It encodes the visual information into structured labels, which can benefit the index and fast retrieval of images in broad practical applications, such as the search engine [24, 27] and the recommendation system [2, 33].

Benefited from the development of deep learning, MLR has achieved remarkable progress in recent years. However, collecting high-quality full annotations becomes very challenging when the label set scales up, which greatly hinders the wide usage of MLR in real scenarios. Recently, researchers explore more feasible solutions for MLR. For example, the full label setting is relaxed with a *partial label* setting in [3, 21], which merely annotates a few labels for each training image. One more extreme setting with solely one *single positive label* is tackled in [8, 16]. These settings can be unified into a common issue of *incomplete labels*, which relieves the burden of the full annotation and considerably reduces the annotation cost. Therefore, it draws increasing attention from both academia and industry.

Compared with the full label setting, the incomplete label setting encounters a dilemma of poor supervision, resulting in severe performance drops for MLR. Existing methods strive to regain supervision from missing labels by exhaustively exploring the *image-to-label* correspondence via semantic-aware modules [4, 21] or loss calibration meth-

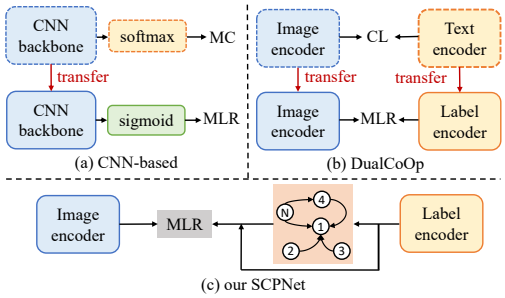

Figure 1. Overview of CNN-based, DualCoOp [26] and our SCPNet. Like DualCoOp, our SCPNet adopts CLIP as the base model. Differently, our SCPNet aims to enhance the MLR with the prior about the *label-to-label* correspondence. MC means multi-class. CL denotes contrastive learning.

*Equal contributions. † Corresponding author.

ods [8, 16, 32]. A convolutional neural network (CNN) pretrained on the ImageNet is usually leveraged to construct the MLR model. Its multi-class softmax layer is often replaced by a multi-label sigmoid layer (Fig. 1 (a)). Such a replacement wipes out prior knowledge about the correspondence between images and labels although it is necessary and inevitable.

Recently, vision-language pretrained models have obtained remarkable success in various vision tasks [26, 34, 35]. Thanks to their large-scale pretraining, the vision-language model, *e.g*., CLIP [22], which is trained with 400 million image-text pairs, can well bridge the visual-textual gap [26], providing rich prior knowledge for the downstream tasks. For the MLR task, Sun *et al*. [26] propose a DualCoOp method, which is the first work to employ the CLIP as the MLR base model. Through dual prompts, DualCoOp directly adopts the text encoder in the CLIP as the multi-label classification head (Fig. 1 (b)), without abandoning the visual-textual prior in the pretrained CLIP.

Despite its effectiveness, DualCoOp is still limited in remedying the deficiency of label supervision, which is desired for the MLR with incomplete labels. Intuitively, it is convenient to reason unknown labels from annotated labels by leveraging the correspondence among labels, *e.g*., tables are likely to appear with chairs, and cars are usually accompanied by roads. Therefore, such a *label-to-label* correspondence can help survive more label supervision and thus benefit MLR with incomplete labels. Besides, although most vision-language models do not encourage the contrastive learning among texts, they are still abundant in the knowledge about the *label-to-label* correspondence because of the large-scale cross-modality training. However, such a valuable prior is rarely explored in the existing state-of-the-art method, *i.e*., DualCoOp [26].

In this paper, we aim to mitigate such deficiency of label supervision for MLR with incomplete labels by leveraging the abundant prior about the *label-to-label* correspondence in the CLIP [22]. We present a structured prior prompter to conveniently derive a structured semantic prior from the CLIP. Then we propose a novel Semantic Correspondence Prompt network (SCPNet) (Fig. 1 (c)), which can prompt the structured label-to-label correspondence with a cross-modality prompter. Our SCPNet also equips a semantic association module to explore high-order relationships among labels with the guidance of the derived structured semantic prior. A prior-enhanced self-supervised learning method is further introduced to comprehensively investigate the valuable prior. As a result, our method can neatly calibrate its predicted semantic distribution while maintaining the self-consistency.

To verify the effectiveness of the proposed method for MLR with incomplete labels, we conduct extensive experiments and analyses on a series of widely used benchmark datasets, *i.e*., MS COCO [19], PASCAL VOC [11], NUS Wide [7], CUB [28] and OpenImages [17]. Experimental results show that our method can significantly outperform state-of-the-art methods on all datasets with a maximal improvement of $6.8\%/3.4\%$ mAP for the single positive label setting and the partial label setting, respectively, well demonstrating its effectiveness and superiority.

Overall, our contributions are four folds.

- We advocate leveraging a structured semantic prior to deal with the deficiency of label supervision for MLR with incomplete labels. To this end, we extract such a prior via a structured prior prompter.
- We present a semantic correspondence prompt Network (SCPNet) based on a cross-modality prompter and a semantic association module. The SCPNet can adequately explore the structured prior knowledge, thus boosting MLR with incomplete labels.
- We design a prior-enhanced self-supervised learning method to further investigate such a structured semantic prior, which can enjoy both distribution refinement and self-consistency.
- Experimental results show that our method can consistently achieve state-of-the-art performance on all benchmark datasets, revealing the significant effectiveness. Thorough analyses also demonstrate the superiority of our method.

2. Related work

Multi-label recognition with full annotations. Multi-label Recognition has long been a hot topic in the computer vision field [1, 21, 30]. A generic method is to learn multiple binary classifiers [8, 16], which usually takes no consideration of the label correlation. Recently, the label-to-label correspondence is established through graph neural networks or transformer structures [6, 29]. These methods heavily rely on the quality of label supervision. However, collecting a large-scale dataset with complete labels is challenging and expensive. In real scenarios, researchers explore much more practical settings with incomplete labels, *i.e*., MLR with partial labels and MLR with a single positive label.

Multi-label recognition with incomplete labels. In the partial label setting, only a few labels need to be annotated for each training image. Durand *et al*. [9] adopt a curriculum learning based model to predict the missing labels during the training procedure. Pu *et al*. [21] and Chen *et al*. [4] transfer predictions of neighboring images via image-image correlation. However, their performance is not guaranteed in more severe scenarios, *i.e*., single positive label setting, in which each image is provided with solely one positive annotation. To tackle the issue of the single positive label,

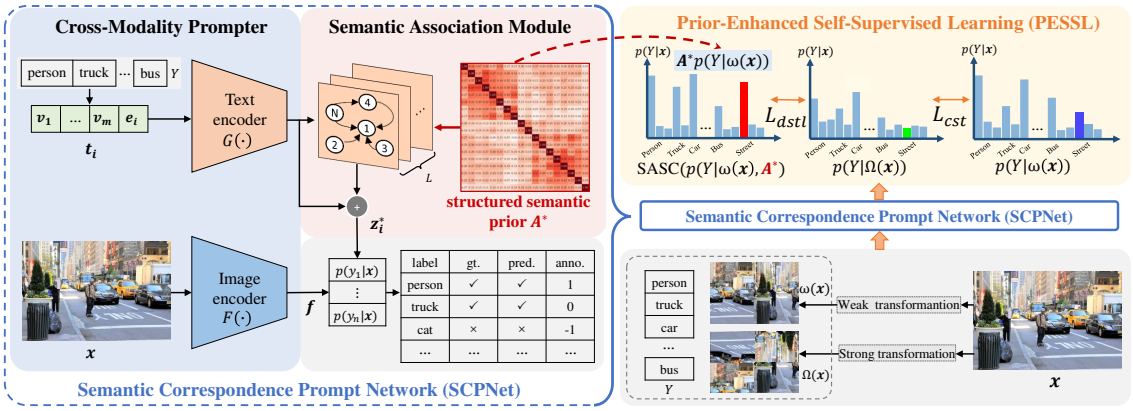

Figure 2. An overview of the proposed method. We design a semantic correspondence prompt network to explore the structured semantic prior for MlR with incomplete labels. A prior-enhanced self-supervised learning strategy is used to enhance such exploration.

Cole *et al*. [8] propose a regularized online loss via a joint optimization of label estimator and image classifier. Zhang *et al*. [32] adopt a label correction process for the probability exceeding a fixed threshold. Kim *et al*. [16] propose to reject or correct the large loss samples during training, which can prevent over-fitting false negative labels. However, different from our solution, they usually independently calibrate the importance of different labels [8, 16, 32], taking no consideration of the semantic correspondence among labels.

Vision-language models in downstream visual tasks. Radford *et al*. [22] exploit the contrastive learning with large-scale image-text pairs, *i.e*., about 400 million pairs, ending up with a powerful vision-language model, *i.e*., CLIP. Such a model shows remarkable generalization capability in downstream visual tasks [22]. Therefore, researchers exhaustively explore how to leverage the abundant vision-language correspondence [12, 14, 23, 26]. Sun *et al*. [26] also employ CLIP for MLR. They present dual prompts, *i.e*., a positive prompt and a negative one, to explore the rich image-to-label correspondence in CLIP. However, different from our motivation, they overlook the rich label-to-label correspondence in CLIP.

3. Methodology

3.1. Structured Prior Prompter

For MLR with full annotations, existing methods can achieve fruitful outcomes by exploring the semantic correspondence between images and labels [6]. However, they require abundant label supervision to obtain accurate label co-occurrence information for the estimation of label relationships. Therefore, in MLR with incomplete labels, the scarce label supervision greatly hinders their capability to explore the semantic correspondence. Benefited from the development of large-scale pretrained embeddings, *e.g*., Glove [20], or models, *e.g*., BERT [15] and CLIP [22], we can easily obtain contextual representations for labels, which can be directly used to derive such a label-to-label correspondence. Such a annotation-free strategy is notably appealing when no adequate label supervision is provided. Furthermore, the abundant correspondence prior in the pretrained model can help associate the annotated label with unknown labels, which promisingly alleviates the deficiency of label supervision. Hence, we introduce a structured prior prompter to explore such a label-to-label correspondence in the pretrained model. Considering the popularity and the remarkable performance in the computer vision community, we choose the vision-language model, *i.e*., CLIP [22], as the target.

Specifically, in the proposed structured prior prompter, for a set of to-be-explored labels $Y = \{y_0, y_1, ..., y_n\}$, we derive the label feature by feeding a prompt template, *i.e*., *a photo of a [CLS]*, into the text encoder of CLIP. We denote the label feature as $\bar{\boldsymbol{z}}_{\boldsymbol{i}}$ for each y_i. Then the correlation prior among labels, denoted as $\boldsymbol{A} = (a_{ij})_{n\times n}$, can be derived as:

$$a_{ij} = \text{sim}(\bar{\boldsymbol{z}}_{\boldsymbol{i}}, \bar{\boldsymbol{z}}_{\boldsymbol{j}}) \tag{1}$$

where $\text{sim}(\cdot,\cdot)$ is the cosine similarity.

For each entry $\boldsymbol{a}_i$, we select the top K elements and set the rest to zero, ending up with a sparse matrix, $\boldsymbol{A'} = (a'_{ij})_{n\times n}$:

$$a'_{ij} = \begin{cases} a_{ij}, & \text{if } j \in \text{topK}(\boldsymbol{a}_i) \\ 0, & \text{if } j \notin \text{topK}(\boldsymbol{a}_i) \end{cases} \tag{2}$$

Following [6], we mitigate the over-smoothness of graph representation by adjusting the sparse graph $\boldsymbol{A'}$ as follows:

$$\bar{a}_{ij} = \begin{cases} (s/\sum_{i\neq j'}^{n} a'_{ij'}) \times a'_{ij}, & \text{if } i \neq j \\ 1-s, & \text{if } i = j \end{cases} \tag{3}$$

where s is a hyper-parameter which determines weights assigned to a node itself and its neighboring nodes. The label

correspondence graph $\mathcal{G}$ can be derived as:

$$\boldsymbol{a}_{ij}^{*} = \frac{\mathrm{I}[\bar{a}_{ij} \neq 0] \exp(\bar{a}_{ij}/\tau^{'})}{\sum_{j} \mathrm{I}[\bar{a}_{ij} \neq 0] \exp(\bar{a}_{ij}/\tau^{'})} \tag{4}$$

where $\tau^{'}$ controls the distribution smoothness and $\mathrm{I}[\cdot]$ is an indicator function. We denote the the adjacency matrix of $\mathcal{G}$ as $\boldsymbol{A^*} = (a_{ij}^{*})_{n \times n}$.

We see that $\boldsymbol{A^*}$ emphasizes the importance of the node itself and weights other nodes according to their relationships (see Eq. (1)). Therefore, the fruitful label correspondence can be encoded in such a structured graph, *i.e.*, $\boldsymbol{A^*}$, providing rich structured semantic prior for MLR models.

3.2. Semantic Correspondence Prompt Network

As shown in Fig. 2, the SCPNet consists of a cross-modality prompter and a semantic association module.

Cross-modality prompter (CMP). Previous works [4, 16, 21] usually employ a convolutional neural network pre-trained on ImageNet, *e.g.*, ResNet50. During fine-tuning in the downstream MLR tasks, the prior knowledge about the image-to-label correspondence is generally discarded due to the semantic shift, *i.e.*, different label sets between the ImageNet and the MLR benchmark datasets. Differently, we aim to take full use of such an image-label prior during model optimization. Similar to [26], we resolve the problem of semantic shift by a cross-modality prompter, based on a vision-language model, *i.e.*, CLIP [22].

Formally, following [35], given a label set, *i.e.*, $Y = \{y_0, y_1, ..., y_n\}$, we introduce m soft prompt tokens to extract its representation. For ease of explanation, we denote the prompt as $\boldsymbol{t}_i = \{\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_m, \boldsymbol{e}_i\}$, where $\boldsymbol{v}$ with a subscript denotes a soft prompt token and $\boldsymbol{e}_i$ is the embedding of y_i. The label feature of y_i, denoted as $\boldsymbol{z_i}$, can be derived by the text encoder of CLIP. For an input image $\boldsymbol{x}$, its visual representation, denoted as $\boldsymbol{f}$, is extracted by the image encoder of CLIP. The process of feature extraction can be computed as follows:

$$\boldsymbol{f} = F(\boldsymbol{x}), \boldsymbol{z_i} = G(\boldsymbol{t_i}), \tag{5}$$

where $F(\cdot)$ and $G(\cdot)$ denote the image encoder and the text encoder in CLIP, respectively.

Semantic association module (SAM). As CMP still lacks capturing the label-to-label correspondence, we further equip a semantic association module to capture high-order relationships among labels. Specifically, with guidance of the structured semantic prior $\boldsymbol{A}^*$ (see Eq. (4)), we utilize L graph convolutional network (GCN) layers to progressively refine the input features $\boldsymbol{H}^0 = \boldsymbol{Z}$, where $\boldsymbol{Z} = \{\boldsymbol{z}_0, \boldsymbol{z}_1, ..., \boldsymbol{z}_n\}$ is a combination of features for Y as in Eq. (5). The l-th GCN layer is updated as follows:

$$\boldsymbol{H}^{l+1} = \rho(\boldsymbol{A}^{*}\boldsymbol{H}^{l}\boldsymbol{W}^{l}), \tag{6}$$

where $\boldsymbol{W}$ with a superscript is a learnable parameter matrix and ρ is a non-linear function. $l \in [0, L)$. The final refined label representations can be obtained through a residual connection, *i.e.*, $\boldsymbol{Z}^{*} = \boldsymbol{H}^{0} + \boldsymbol{H}^{L}$. The likelihood $p(y_i|\boldsymbol{x})$ can be computed as:

$$p(y_i|\boldsymbol{x}) = \sigma(\mathrm{sim}(\boldsymbol{f}, \boldsymbol{z_i^{*}})/\tau), \tag{7}$$

where $\boldsymbol{z_i}^{*}$ denotes the refined feature for label y_i.

Benefited from the GCN, the structured label-to-label correspondence in CLIP, which is represented by $\boldsymbol{A}^{*}$, can be progressively refined in the label representation. Therefore, during the semantic matching between the image feature and the label feature, *i.e.*, Eq. (7), labels with high correlations will obtain similar likelihoods, enabling a subtle semantic association.

3.3. Prior-Enhanced Self-Supervised Learning

The proposed prior-enhanced self-supervised learning strategy, dubbed PESSL, aims to make full use of the structured semantic correspondence prior. We endow the proposed PESSL with a self-supervised consistency loss and a self-distillation objective that is boosted by a structure-aware semantic calibration strategy.

Structure-aware semantic calibration. Intuitively, if two labels are semantically correlated, they may be observed in one image. For MLR, such a correspondence can help decide potential semantic labels for an input image, given its predictions. Therefore, we formulate the likelihood of $p(y_i|\boldsymbol{x})$ as a weighted combination of likelihoods for correlated neighboring labels of y_i:

$$p^{*}(y_i|\boldsymbol{x}) = \sum\nolimits_{y_j \in \mathcal{N}(y_i)} w(i,j) \times p(y_j|\boldsymbol{x}) \tag{8}$$

Here, $w(i,j)$ is a correlation weight indicating the relationship between y_i and y_j. $\mathcal{N}(y_i)$ denotes a correlated neighboring set of labels corresponding to y_i.

For ease of explanation, we introduce a correlation matrix $\mathbb{W}$ to represent the whole correlation among labels, *i.e.*, $\mathbb{W} = (w(i,j))_{n \times n}$. We then customize the whole process as a function parameterized by $\mathbb{W} \in R^{n \times n}$ and the distribution over Y given the input $\boldsymbol{x}$, *i.e.*, $\boldsymbol{p}(Y|\boldsymbol{x}) \in R^{n \times 1}$:

$$\mathrm{SASC}(\boldsymbol{p}(Y|\boldsymbol{x}), \mathbb{W}) = \mathbb{W}\boldsymbol{p}(Y|\boldsymbol{x}) \tag{9}$$

Prior-enhanced learning. Existing loss correction methods individually reweight each label, without taking into consideration the correspondence among labels. Here, we propose to follow the self-supervised learning principle [25, 31] and introduce a self-distillation learning strategy to benefit the MLR model from the structured semantic correspondence among labels.

Specifically, we derive two different versions for the input image $\boldsymbol{x}$ with one weak transformation $\omega(\cdot)$ and

one strong transformation $\Omega(\cdot)$, respectively. Their corresponding semantic distributions, denoted as $\boldsymbol{p}(y|\omega(\boldsymbol{x}))$ and $\boldsymbol{p}(y|\Omega(\boldsymbol{x}))$, respectively, can be derived by Eq. (7). Then we use a consistency loss to encourage them to be consistent. Different from [31], which simply regularizes the model with the most confident label, we construct a set of confident labels $\mathcal{O}(\boldsymbol{x})$ with the top highest probability larger than a threshold $\mathcal{T}$ in $\boldsymbol{p}(y|\omega(\boldsymbol{x}))$, *i.e.*, $\mathcal{O}(\boldsymbol{x}) = \{c|c \in \text{topK}(\boldsymbol{p}(y|\omega(\boldsymbol{x}))) \wedge p(c|\omega(\boldsymbol{x})) > \mathcal{T}(c)\}$. A dynamic threshold strategy is performed for each label, as [31]. The consistency loss is then derived by:

$$\begin{aligned}\mathcal{L}_{cst} = &-\sum\nolimits_{c\in\mathcal{O}(\boldsymbol{x})}^{Y} \log p(c|\Omega(\boldsymbol{x})) \\ &-\sum\nolimits_{c\notin\mathcal{O}(\boldsymbol{x})}^{Y} \log(1-p(c|\Omega(\boldsymbol{x})))\end{aligned} \tag{10}$$

We calibrate the distribution of the weak-transformed image, *i.e.*, $\boldsymbol{p}(y|\omega(\boldsymbol{x}))$, by using the SASC function (see Eq. (9)):

$$\boldsymbol{p}^*(y|\omega(\boldsymbol{x})) = \text{SASC}(\boldsymbol{p}(y|\omega(\boldsymbol{x})), \boldsymbol{A}^*) \tag{11}$$

where $\boldsymbol{A}^*$ represents the structured semantic prior, derived by Eq. (4). Considering that compared with the weak-transformed image, the strong-transformed image is usually more difficult to learn. Therefore, we employ a self-distillation objective to optimize the distribution of the strong-transformed image $\Omega(\boldsymbol{x})$ with the guidance of the calibrated semantic distribution via the KL-divergence:

$$\mathcal{L}_{dstl} = -\sum_{c}^{Y}\left(q_c^w \log\frac{q_c^s}{q_c^w} + (1-q_c^w)\log\frac{1-q_c^s}{1-q_c^w}\right) \tag{12}$$

where $q_c^w = p^*(c|\omega(\boldsymbol{x}))$ and $q_c^s = p(c|\Omega(\boldsymbol{x}))$.

Overall Objective. Finally, we formulate the prior-enhanced self-supervised learning as a combination of the consistency objective and the self-distillation objective:

$$\mathcal{L}_{pessl} = \lambda_{cst}\mathcal{L}_{cst} + \lambda_{dstl}\mathcal{L}_{dstl} \tag{13}$$

3.4. Network Optimization

During training, we adopt a multi-label classification objective over the predicted likelihood, *i.e.*, $p(y_i|\boldsymbol{x})$ in Eq. (7), to optimize our SCPNet, denoted as $\mathcal{L}_{cls}$. We follow [32] to design $\mathcal{L}_{cls}$. The overall objective for the network optimization is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{pessl} \tag{14}$$

4. Experiment

4.1. Experiment Settings

Datasets. We conduct extensive experiments on several standard benchmarks for MLR with incomplete labels, including the single positive label setting and the partial label setting. For the single positive label setting, following [16, 32], we use MS-COCO (COCO) [19], PASCAL VOC (VOC) [11], NUSWIDE (NUS) [7], and CUB [28]. For the partial label learning, we adopt MS-COCO (COCO) [19], PASCAL VOC 2007 (VOC2007) [10] and Visual Genome (VG-200) [18], as [4, 21]. We leave details of benchmark datasets in the supplementary due to the space limit.

Implementation details. We leverage published CLIP weights[1] to initialize MLR models. To fairly compare the proposed method with others, we adopt the ResNet50-based CLIP and the Resnet101-based CLIP for the single positive label and the partial label, respectively. During training, we tune the image encoder and fix the text encoder of CLIP. More details are provided in the supplementary.

Evaluation. By default, we employ the mean average precision (mAP) as the evaluation metric, following previous works [5, 16, 32]. For the single positive label setting, we perform two different setups, *i.e.*, the LargeLoss setup [16] and the SPLC setup [32], which are common in the community. We leave the details in the supplementary due to the space limit. For the partial label setting, following [21], we randomly maintain partial labels for the training set with a ratio ranging from 10% to 90%. Apart from performance on all ratios, we also report the average result.

4.2. Comparisons with State-of-the-Arts

MLR with single positive labels. We report the model performance on both the LargeLoss setup [16] and the SPLC setup [32]. To better reveal the effectiveness of the proposed method, we also report the average performance for both setups. As shown in Tab. 1, for both setups, our method can significantly outperform existing methods on all benchmark datasets, achieving state-of-the-art performance. Specifically, in the LargeLoss setup, the proposed SCPNet can obtain a maximal performance improvement of 4.7% (NUS). As a whole, our method can accomplish an overall performance improvement of 3.6%. In the SPLC setup, the maximal performance improvement achieved by our method can reach 6.8% (NUS). As a result, our method can accomplish 4.7% improvement on average.

MLR with partial labels. As shown in Tab. 2, our results also consistently surpass existing state-of-the-art methods on all benchmark datasets, especially on the COCO and VG-200. Compared with DualCoOp [26] which also leverages CLIP to build MLR models, the proposed method can obtain an improvement of 1.9% mAP on the MS COCO. With a frozen image encoder during training as DualCoOp, our method, denoted as SCPNet (ours)*, still enjoys superior performance to DualCoOp. On the VOC2007, our method obtains comparable performance with 0.3% improvement. However, under small ratios, our method shows its superiority to DualCoOp, *e.g.*, 0.8% improvement with a ratio of 10%. On the VG-200, compared with SARB [21]

[1] https://github.com/openai/CLIP

Table 1. Comparison with the state-of-the-art methods for MLR with the single positive label (%).

Method	LargeLoss setup [16]					SPLC setup [32]				
	COCO	VOC	NUS	CUB	Avg.	COCO	VOC	NUS	CUB	Avg.
LSAN [8]	69.2	86.7	50.5	17.9	56.1	70.5	87.2	52.5	18.9	57.3
ROLE [8]	69.0	88.2	51.0	16.8	56.3	70.9	89.0	50.6	20.4	57.7
LargeLoss [16]	71.6	89.3	49.6	21.8	58.1	-	-	-	-	-
Hill [32]	-	-	-	-	-	73.2	87.8	55.0	18.8	58.7
SPLC [32]	72.0	87.7	49.8	18.0	56.9	73.2	88.1	55.2	20.0	59.1
SCPNet (ours)	**75.4**	**90.1**	**55.7**	**25.4**	**61.7**	**76.4**	**91.2**	**62.0**	**25.7**	**63.8**

Table 2. Comparison with the state-of-the-art methods for MLR with partial labels (%).

Datasets	Method	10%	20%	30%	40%	50%	60%	70%	80%	90%	Avg.
COCO	SSGRL [5]	62.5	70.5	73.2	74.5	76.3	76.5	77.1	77.9	78.4	74.1
	GCN-ML [6]	63.8	70.9	72.8	74.0	76.7	77.1	77.3	78.3	78.6	74.4
	SST [4]	68.1	73.5	75.9	77.3	78.1	78.9	79.2	79.6	79.9	76.7
	SARB [21]	71.2	75.0	77.1	78.3	78.9	79.6	79.8	80.5	80.5	77.9
	DualCoOp [26]	78.7	80.9	81.7	82.0	82.5	82.7	82.8	83.0	83.1	81.9
	SCPNet (ours)*	**80.3**	**82.2**	**82.8**	83.4	83.8	83.9	84.0	84.1	84.2	83.2
	SCPNet (ours)	79.1	82.1	**82.8**	**83.9**	**84.5**	**84.9**	**85.4**	**85.7**	**85.9**	**83.8**
VOC2007	SSGRL [5]	77.7	87.6	89.9	90.7	91.4	91.8	91.9	92.2	92.2	89.5
	GCN-ML [6]	74.5	87.4	89.7	90.7	91.0	91.3	91.5	91.8	92.0	88.9
	SST [4]	81.5	89.0	90.3	91.0	91.6	92.0	92.5	92.6	92.7	90.4
	SARB [21]	83.5	88.6	90.7	91.4	91.9	92.2	92.6	92.8	92.9	90.7
	DualCoOp [26]	90.3	92.2	92.8	93.3	93.6	93.9	94.0	94.1	94.2	93.2
	SCPNet (ours)	**91.1**	**92.8**	**93.5**	**93.6**	**93.8**	**94.0**	**94.1**	**94.2**	**94.3**	**93.5**
VG-200	SSGRL [5]	34.6	37.3	39.2	40.1	40.4	41.0	41.3	41.6	42.1	39.7
	GCN-ML [6]	32.0	37.8	38.8	39.1	39.6	40.0	41.9	42.3	42.5	39.3
	SST [4]	38.8	39.4	41.1	41.8	42.7	42.9	43.0	43.2	43.5	41.8
	SARB [21]	41.4	44.0	44.8	45.5	46.6	47.5	47.8	48.0	48.2	46.0
	SCPNet (ours)	**43.8**	**46.4**	**48.2**	**49.6**	**50.4**	**50.9**	**51.3**	**51.6**	**52.0**	**49.4**

which enhances the MLR models with a structure-aware algorithm, our SCPNet can significantly outperform it with an average performance improvement of 3.4%.

These experimental results show that our method can consistently obtain superior performance in different setups for MLR with incomplete labels, well demonstrating the effectiveness. To verify the generalization of the proposed method, we also investigate the effectiveness in the few-shot partial label setting and the real partial label scenario. We leave them in the supplementary due to the space limit.

4.3. Ablation Study

In order to analyze the effectiveness of each component, we conduct the ablation study on both the single positive label and the partial label settings. All results are shown in Tab. 3. We also introduce a model that directly employs $\mathcal{L}_{cls}$ to optimize a ResNet-based MLR model, as the baseline. As shown in Tab. 3, each component can obtain consistent performance improvement in all datasets. Specifically, compared with the baseline model, our CMP can obtain an average performance of 1.99% mAP, indicating the superiority of prompting a cross-modality vision-language model. Augmented by SAM, our method can bring 0.69% mAP improvement. Such improvements can be attributed to the explicit semantic correspondence among labels captured by the proposed SAM component. Besides, the consistency learning, *i.e.*, $\mathcal{L}_{cst}$, and the self-distillation objective, *i.e.*, $\mathcal{L}_{dstl}$, can lead to 1.25% and 1.56% performance improvement, respectively. The overall improvement for the proposed PESSL can reach 2.09%, well demonstrating the strength of incorporating the structured semantic prior during model optimization. Finally, our proposed SCPNet can significantly outperform the baseline model with 4.77% mAP improvement on average, well demonstrating the effectiveness and the superiority of the proposed method.

4.4. Model Analysis

Here, we perform comprehensive inspections for the proposed method. All experiments are conducted in the single positive label setting on the MS COCO dataset, by default. Due to the space limit, we provide more analyses in the supplementary material.

Table 3. Effect of different modules in the proposed SCPNet method for both the single positive label setting and the partial label setting (%). An average of all metrics is also reported.

Model	CMP	SAM	PESSL		Single Positive Label				Partial Label			Avg.
			$\mathcal{L}_{cst}$	$\mathcal{L}_{dstl}$	COCO	VOC	NUS	CUB	COCO	VOC2007	VG-200	
Baseline					73.18	88.07	55.18	19.99	77.41	88.32	46.39	64.08
SCPNet	✓				74.36	88.46	60.66	21.42	80.90	89.16	47.55	66.07
	✓	✓			75.12	89.09	61.08	21.66	82.12	90.16	48.11	66.76
	✓	✓	✓		75.70	90.92	61.75	23.67	82.85	92.50	48.70	68.01
	✓	✓		✓	75.84	90.92	61.56	24.51	83.35	93.21	48.83	68.32
	✓	✓	✓	✓	**76.42**	**91.16**	**62.04**	**25.71**	**83.76**	**93.49**	**49.36**	**68.85**

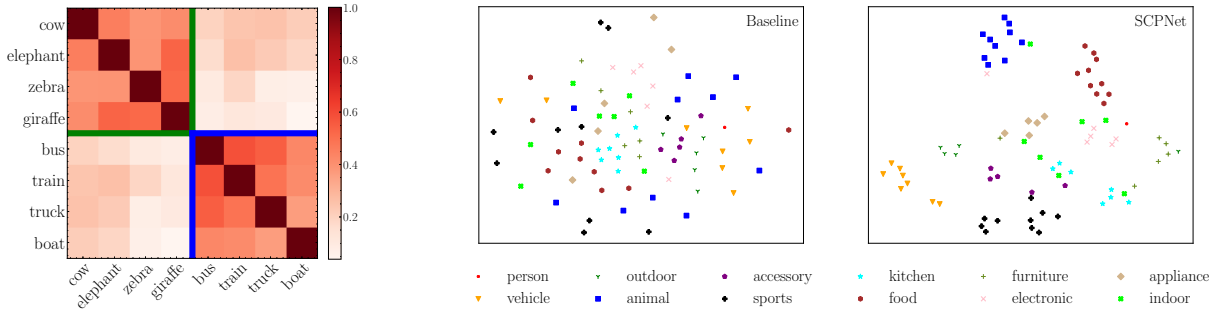

Figure 3. The structured semantic prior (left) and the learnt label representation (middle: in the baseline, right: in our SCPNet).

Table 4. Analysis on the correlation graph.

SAM	PESSL	mAP (%)
Static	Static	**76.42**
	Dynamic	76.05
	No	75.83
Dynamic	Static	76.08
	Dynamic	75.84

Table 5. Analysis on the prior extraction (%).

Prior	Dynamic	Image	Glove	BERT	CLIP
mAP	75.84	75.67	76.15	76.16	**76.42**

Correlation graph construction. We verify the positive effect of the prior used in the correlation graph construction for both SAM and PESSL. To achieve this goal, we discuss two kinds of correlation graph: 1) a static one derived from the pretrained CLIP model (see Eq. (4)), which captures the structured semantic prior, and 2) a dynamic one achieved by the learnable CMP, *i.e.*, constructing the adjacency matrix with label features $\boldsymbol{z_i}$ computed by Eq. (5). We also report PESSL without the prior, denoted as "No". As illustrated in Tab. 4, we can observe that our method can obtain the optimal performance by using the static correlation graph for both SAM and PESSL. Besides, the static graph can substantially achieve better results than the dynamic one in both components, revealing the advantage of the structured semantic prior. We claim that in the MLR with incomplete labels, the challenge of insufficient label supervision makes the dynamic graph sub-optimal, thus inferior to the static one. By comparing PESSL with the prior (Row 2) and the one without the prior (Row 4), we can find that the latter achieves inferior performance, which can demonstrate the benefit of the proposed prior, *i.e.*, $\boldsymbol{A}^*$ in Eq. (4).

Prior knowledge extraction. We further investigate the advantage of the proposed structured semantic prior extracted by CLIP with three other types of prior knowledge as competitors. For a given label, 1) "Image" averages all image features corresponding to it; 2) "Glove" represents its feature by pretrained Glove word embeddings [20]; and 3) "BERT" extracts the label feature by the prompt learning as ours. We also report the result of dynamic label-to-label correspondence as the baseline. As shown in Tab. 5, compared with Dynamic, except Image, Glove, BERT and CLIP show consistent advantage because of their superior ability to capture the label-to-label correspondence. Besides, our CLIP achieves the best performance, which reveals that the CLIP-based structured prior is more matched with the CLIP-based MLR model due to their consistent knowledge.

Generalization on the CNN-based architecture. To show the generalization of the proposed prior-enhanced method, we transfer our design principles to a vanilla

Table 6. Prior for MLR models with the ImageNet-based ResNet.

Image Encoder	Label Encoder	mAP (%)
ResNet	sigmoid	73.18
ResNet	Ours	74.72
Ours	Ours	**76.42**

Table 7. Analysis on the number of GCN Layer, *i.e.*, L (%).

L	2	3	4
mAP	75.88	**76.42**	76.22

Table 8. Analysis on λ_{cst} and λ_{dstl} (%).

λ_{cst}	0	1/16	1/8	1/4	1/8		
λ_{dstl}	0				1/8	2/8	3/8
mAP	75.12	75.56	**75.70**	75.13	**76.42**	76.40	76.17

ResNet-based MLR model with a sigmoid layer as the label encoder. We analyze the impact of replacing the sigmoid layer with ours. As shown in Tab. 6, such modification can result in a performance gain of 1.54%, demonstrating the good generalization ability of the proposed method in the CNN-based architecture.

Analysis on hyper-parameters. As shown in Tab. 7 and Tab. 8, the best value of L, λ_{cst} and λ_{dstl} is at $L = 3$, $\lambda_{cst} = 1/8$, and $\lambda_{dstl} = 1/8$, respectively. More analyses can be found in the supplementary material.

4.5. More Insightful Analysis

To provide more insights about the effectiveness of the proposed method, we conduct visualization analyses on the structured semantic prior and the label representation in the latent feature space. First, we present the structured semantic prior about the label-to-label correspondence by visualizing the adjacency matrix, *i.e.*, $\boldsymbol{A}^*$ in Eq. (4) on MS COCO. For ease of explanation, we select two categories, *i.e.*, animal and vehicle, and investigate the label correspondence among labels associated with them. As shown in Fig. 3, the used structured semantic prior can successfully convey the similarity among labels although the CLIP is not encouraged in the contrastive learning on the text. Second, we visualize the label features in the baseline (weights in the sigmoid layer) and our SCPNet (output of the SAM, denoted as $\boldsymbol{z_i}^*$ in Eq. (7)). We can observe that labels belonging to the same category are more well-aligned together in our SCPNet, compared with those in the baseline model. This result indicates that our SCPNet can reasonably derive more discriminative label representations due to the appliance of the structured semantic prior.

To verify the effect of the proposed structured semantic prior on the issue of insufficient label supervision, we introduce a competitor model, *i.e.*, CMP+$\mathcal{L}_{cst}$, which wipes out components involving the prior, *i.e.*, $\boldsymbol{A}^*$ in Eq. (4). We keep track of the precision of model predictions on the training set and the mAP result on the test set after each training epoch. For CMP+$\mathcal{L}_{cst}$, we also visualize the precision over its calibrated predictions by the SASC$(\cdot)$ function. As illustrated in Fig. 4 (left), compared with CMP+$\mathcal{L}_{cst}$, both SASC(CMP+$\mathcal{L}_{cst}$) and our SCPNet can obtain consistent improvements in terms of the label prediction precision. It indicates that the quality of label supervision can be promoted under the guidance of the proposed prior, thus benefiting the performance on the test set (see Fig. 4 (right)).

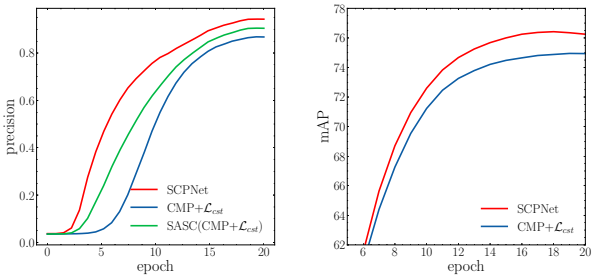

Figure 4. The precision on the training set (left) and the mAP on the test set (right).

5. Conclusion

In this paper, we drive a structured semantic prior about the label-to-label correspondence from the vision-language model, *i.e.*, CLIP [22]. To mitigate the deficiency of label supervision for MLR with incomplete labels, we introduce a semantic correspondence prompt network, dubbed SCPNet, which can explore such a structured semantic prior. It constructs a cross-modality prompter to leverage the explicit image-to-label correspondence in the CLIP. A semantic association module is equipped to associate related labels with the help of such a meaningful structured semantic prior. Furthermore, we propose a prior-enhanced self-supervised learning method for network optimization. Experimental results on a series of benchmark datasets for MLR with incomplete labels show that our method can achieve state-of-the-art performance on both the partial label setting and the single positive label setting, well demonstrating its effectiveness and superiority. In the future, we will further study how to generalize our method to tackle other practical problems, *e.g.*, the domain gap.

Acknowledgement. This work was supported by "Pioneer" and "Leading Goose" R&D Program of Zhejiang (No. 2023C01038), National Natural Science Foundation of China (Nos. 62271281, 61773301), Zhejiang Provincial Natural Science Foundation of China under Grant (No. LDT23F01013F01), China Postdoctoral Science Foundation (No. BX2021161) and Shanxi Innovation Team Project (No. 2018TD-012).

References

[1] Emanuel Ben-Baruch, Tal Ridnik, Itamar Friedman, Avi Ben-Cohen, Nadav Zamir, Asaf Noy, and Lihi Zelnik-Manor. Multi-label classification with partial annotations using class-aware selective loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4764–4772, 2022. 2

[2] Dolly Carrillo, Vivian F López, and María N Moreno. Multi-label classification for recommender systems. *Trends in Practical Applications of Agents and Multiagent Systems*, pages 181–188, 2013. 1

[3] Tianshui Chen, Liang Lin, Xiaolu Hui, Riquan Chen, and Hefeng Wu. Knowledge-guided multi-label few-shot learning for general image recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1

[4] Tianshui Chen, Tao Pu, Hefeng Wu, Yuan Xie, and Liang Lin. Structured semantic transfer for multi-label recognition with partial labels. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 339–346, 2022. 2, 4, 5, 6

[5] Tianshui Chen, Muxin Xu, Xiaolu Hui, Hefeng Wu, and Liang Lin. Learning semantic-specific graph representation for multi-label image recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 522–531, 2019. 1, 5, 6

[6] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5177–5186, 2019. 2, 3, 6

[7] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, pages 1–9, 2009. 2, 5

[8] Elijah Cole, Oisin Mac Aodha, Titouan Lorieul, Pietro Perona, Dan Morris, and Nebojsa Jojic. Multi-label learning from single positive labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 933–942, 2021. 1, 2, 3, 6

[9] Thibaut Durand, Nazanin Mehrasa, and Greg Mori. Learning a deep convnet for multi-label classification with partial labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 647–657, 2019. 2

[10] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 5

[11] Mark Everingham and John Winn. The pascal visual object classes challenge 2012 (voc2012) development kit. *Pattern Anal. Stat. Model. Comput. Learn., Tech. Rep*, 2007:1–45, 2012. 2, 5

[12] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*, 2020. 3

[13] Dat Huynh and Ehsan Elhamifar. Interactive multi-label cnn learning with partial labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9423–9432, 2020.

[14] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020. 3

[15] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019. 3

[16] Youngwook Kim, Jae Myung Kim, Zeynep Akata, and Jungwoo Lee. Large loss matters in weakly supervised multi-label classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14156–14165, 2022. 1, 2, 3, 4, 5, 6

[17] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2(3):18, 2017. 2

[18] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 5

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 5

[20] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014. 3, 7

[21] Tao Pu, Tianshui Chen, Hefeng Wu, and Liang Lin. Semantic-aware representation blending for multi-label image recognition with partial labels. *arXiv preprint arXiv:2203.02172*, 2022. 1, 2, 4, 5, 6

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 4, 8

[23] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*, 2020. 3

[24] Josef Sivic and Andrew Zisserman. Video google: Efficient visual search of videos. In *Toward category-level object recognition*, pages 127–144. Springer, 2006. 1

[25] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying

semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020. 4

[26] Ximeng Sun, Ping Hu, and Kate Saenko. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. *arXiv preprint arXiv:2206.09541*, 2022. 1, 2, 3, 4, 5, 6

[27] Ivona Tautkute, Tomasz Trzciński, Aleksander P Skorupa, Łukasz Brocki, and Krzysztof Marasek. Deepstyle: Multimodal search engine for fashion and interior design. *IEEE Access*, 7:84613–84628, 2019. 1

[28] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 2, 5

[29] Ya Wang, Dongliang He, Fu Li, Xiang Long, Zhichao Zhou, Jinwen Ma, and Shilei Wen. Multi-label classification with label graph superimposing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12265–12272, 2020. 1, 2

[30] Vacit Oguz Yazici, Abel Gonzalez-Garcia, Arnau Ramisa, Bartlomiej Twardowski, and Joost van de Weijer. Orderless recurrent models for multi-label classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13440–13449, 2020. 1, 2

[31] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34:18408–18419, 2021. 4, 5

[32] Youcai Zhang, Yuhao Cheng, Xinyu Huang, Fei Wen, Rui Feng, Yaqian Li, and Yandong Guo. Simple and robust loss design for multi-label learning with missing labels. *arXiv preprint arXiv:2112.07368*, 2021. 2, 3, 5, 6

[33] Yong Zheng, Bamshad Mobasher, and Robin Burke. Context recommendation using multi-label classification. In *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 2, pages 288–295. IEEE, 2014. 1

[34] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 2

[35] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 2, 4